Precise Machine Learning

Tae-Geun, Kim

***M.S.**: Dept of Physics, Yonsei University*
***B.S.**: Dept of Astronomy, Yonsei University*

Abstract: 이제 머신러닝은 거의 대부분의 이공계인들 (뿐만 아니라 문과생들에게도) 필수적인 도구가 되었습니다. arXiv만 봐도 매일 머신러닝 관련 논문이 분야를 가리지 않고 나오며 페이스북에는 수 많은 머신러닝 강좌 홍보글들이 쏟아지고 있습니다. 개발자들 사이에서도 이제 머신러닝은 필수적인 기술 스택이 되었습니다. 하지만 안타깝게도 제대로 된 머신러닝 강의는 그리 많지 않습니다. 머신러닝에 접근하는 방법은 크게 세 가지로 나눌 수 있습니다. 하나는 Developer's route 로 간단히 사람들이 만들어 놓은 Framework 들을 이용해 연구에 활용하는 것입니다. 이 방법은 누구나 쉽게 접근할 수 있다는 장점이 있지만, 여기에 안주하다간 간단한 ML 논문 하나 읽기도 벅찰 것입니다. 두 번째는 Statistician's route 로 충분히 통계적 Background를 쌓은 후 Machine Learning을 이해하는 것입니다. 공부하는데 시간이 꽤 걸리지만 습득한 통계적 기술들은 모든 곳에 유용하게 쓰일 수 있습니다. 더불어 계속 쏟아져 나오는 ML의 신 기술들을 받아들이기도 쉽습니다. 마지막은 Mathematician's route 입니다. 본래 확률론은 수학의 영역입니다. 따라서 통계학의 본질에는 항상 수학이 빠질 수 없습니다. 비유를 하자면 통계학이 ML의 Front-end라면 수학은 ML의 Back-end인 셈이죠. 딱히 수학을 배운다고 실용적인 ML에 도움이 되는 것은 아닙니다만 통계에서 받아들이고 넘어가던 것에 대해 당위성을 얻게 됩니다. 본질적인 이해는 활용에도 도움되는 것은 당연하니 결과적으로는 좋은 영향을 미칠 것입니다. 이 문서는 마지막 방법인 Mathematician's route를 차용하여 머신러닝을 설명할 것입니다.

Contents

1 Introduction 1

2 Measure Theory 2

1. Introduction

우리는 자연을 관측하여 그것을 어떠한 논리를 이용하여 정해진 모델로 분류합니다. 이를 수학으로 표현하면 다음과 같습니다.

- Observation: $x \in \mathbb{R}^d$
- Class: $y \in \{1, 2, \cdots, M\}$
- Classifier: $g : \mathbb{R}^d \to \{1, 2, \cdots, M\}$
- Error: When $g(x) \neq y$

우리의 목적은 관측값을 모델로 분류하는 분류기인 g를 찾는 것입니다. 당연하게도 모든 상황에 통용되는 g를 찾을 수 있으면 좋겠지만 안타깝게도 그런 g는 존재하지 않습니다. 하나의 관측 값이 항상 하나의 모델에만 대응되는 것은 아니기 때문이죠. 따라서 분류기에는 항상 error가 존재하기 마련입니다. 즉, 우리의 관측은 비결정적(Undeterministic)이고 이를 표현하기 위해서는 데카르트의 결정적 수학을 넘어서야 합니다. 우리는 이러한 이론을 **확률론**(Probabilistic Theory)이라 부릅니다.

확률론을 사용하기 위하여 단순히 일차원적 변수를 넘어서 하나의 쌍을 메인 변수로 볼 것입니다. 그리고 이제 Error의 정도를 확률을 사용하여 명시할 수 있습니다. 이를 역시 수학으로 기술하면 다음과 같습니다.

- Pair: $(X,Y) \in \mathbb{R}^d \times \{1,\cdots,M\}$
- Probability of Error: $L(g) = \mathbf{P}\{g(X) \neq Y\}$

그리고 이를 활용해 가장 좋은 (완벽하진 않습니다.) 분류기를 정의할 수도 있습니다.

$$g^* = \underset{g:\mathbb{R}^d \to \{1,\cdots,M\}}{\arg\min} L(g) \tag{1}$$

이론 상 완벽하지만, 안타깝게도 우리에겐 (X,Y)의 분포나 g의 값이 주어져 있지 않습니다. 따라서 당연하게도 가장 좋은 분류기 조차 구할 수 없죠. 그렇다면 포기해야 할까요? 다행히 인류는 지금 껏 아주 많은 데이터를 생산했고 우리의 선조들은 그것을 이미 분류해놓았습니다. 우리는 지금까지 누적된 훌륭한 데이터 셋을 이용할 것입니다. 이미 분류되어 있는 데이터 셋들을 이용하면 분류기를 만들 수 있습니다. 또한, 좀 더 간편하게 데이터를 해석하기 위하여 *independent identically distributed*(*i.i.d*) 가정을 이용할 것입니다. 이는 비록 엄청 강력한 가정이긴 하나, 수 많은 연구에서 그리 큰 차이를 만들지 않는다는 결론을 내놓은 바 있습니다. 이제 이를 다시 수학으로 기술하여 봅시다.

- Pre-Classified Data: $\{(X_i,Y_i)\}_{i=1}^n$
- Classifier (Trained): $g_n : \mathbb{R}^d \times \left(\mathbb{R}^d \times \{1,\cdots,M\}\right)^n \to \{1,\cdots,M\}$
- Conditional Probability of Error: $L_n = L(g_n) = \mathbf{P}\{g_n(X;\{(X_i,Y_i)\}_{i=1}^n) \neq Y | \{(X_i,Y_i)\}_{i=1}^n\}$
- Average of L_n: $\mathbf{E}L_n$

위에서도 한 번 언급했지만 우리는 이제 분류기를 데이터 셋들로 훈련시킬 수 있습니다. 따라서 g_n의 차원이 상당히 복잡하죠. 우리는 어디까지나 통계적으로 접근할 것이므로 개개인의 Error는 중요하지 않고 평균이 상당히 중요해집니다. 따라서 $\mathbf{E}L_n$을 사용할 것이고 이것으로 Classifier의 성능을 평가할 것입니다.

이제 기초에 대한 수학적 서술은 대강 끝났습니다. 우리의 목표는 $\mathbf{E}L_n$을 최소화 시키는 것입니다. 이를 위해 우리는 확률론을 사용할텐데 확률론을 공부하기 위해서는 필수로 공부해야 하는 학문이 있습니다. 바로 **측도론**(Measure Theory)입니다.

2. Measure Theory

측도론이란, 아주 간단히 말하면 집합의 크기를 측정하기 위한 학문입니다. 우리가 보통 확률을 처음 접할 때, 다음과 같은 정의를 본 적이 있을 겁니다.

Example 1 - High School Probability.

Probability of occurrence of events A in sample space Ω is

$$P(A) = \frac{n(A)}{n(\Omega)}$$

위의 정의를 이용하면 아주 간단명료하게 확률을 구할 수 있습니다. 주사위를 던졌을 때, 1이 나올 확률은 $\{1,2,3,4,5,6\}$을 전체집합으로 설정하면 전체 6개 중에 1은 1개이기에 확률은 $\frac{1}{6}$이 됩니다. 하지만 다음의 경우에는 난처해집니다.

Example 2 - Common High School Problem.

정사각형 과녁 안에 원 모양 과녁이 네 변에 모두 내접하여 있을 때, 화살이 원 안에 명중할 확률을 구하여라.

간단한 확률 지식이 있는 사람이라면 이 문제를 $\frac{\text{원의 넓이}}{\text{전체 정사각형의 넓이}}$로 접근하여 $\frac{\pi}{4}$임을 알아낼 수 있을 것입니다. 그런데 아까 분명 확률을 사건의 경우의 수를 전체 경우의 수로 나누어 구한다고 했는데 여기서 전체 경우의 수는 얼마일까요? 하다 못해 원의 경우의 수는 어떻게 구해야 할까요? 우리는 지금까지 아무런 의심 없이 구해왔습니다만, 이제부터 의심을 가져야 합니다. 확률은 특정한 경우에는 경우의 수로 구해질 수 있습니다만, 집합이 무한해지는 순간 우리가 정의한 확률은 아무 의미 없어 집니다. 따라서 우리는 확률을 엄밀하게 정의해야 할 필요가 있습니다. 그리고 그러기 위해서는 먼저 집합의 크기를 정의해야 하는 것이 우선입니다. 따라서 측도론이 필요한 것이죠.

필요성은 알았으나 측도론은 상당히 추상적인 학문이라 곧바로 집합의 크기를 어떻다 정의내릴 수 없습니다. 따라서 크기를 어떻게 정의하는지는 모르지만 측정 가능성에 대해 먼저 논해봅시다. 어떤 집합 U,V가 크기를 잴 수 있는 집합(가측집합; Measurable Set)이라 해봅시다. 그렇다면 직관적으로 다음의 사실들을 받아 들일 수 있습니다.

- $U \cup V$ is measurable
- $U \cap V$ is measurable
- U^c, V^c is measurable

하지만 수학자들은 이런 모호한 직관의 정렬을 좋아하지 않습니다. 이런 규칙들을 모아 하나의 우아한 규칙으로 정의내리죠. 수학에서 이러한 규칙을 우리는 대수(Algebra)라고 부릅니다.

Definition 1 - σ-algebra.

Let S be a set, and let $\mathscr{F}$ be a family of subsets of S. $\mathscr{F}$ is called a σ-algebra if
i) $\emptyset \in \mathscr{F}$
ii) $A \in \mathscr{F}$ implies $A^c \in \mathscr{F}$
iii) $A_1, A_2, \cdots \in \mathscr{F}$ implies $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$

정의는 무척 복잡해보이지만, 사실 의미는 간단합니다. 대수는 일종의 규칙을 담아 놓은 집합입니다. 예를 들어 위상수학에서의 Topology는 모든 열린 집합들의 집합이며 여기서 말하는 σ-algebra(시그마대수)는 모든 가측집합들의 집합입니다. 공집합은 너무나 당연하게 그 크기를 0으로 측정할 수 있을 테고 어떤 집합 A가 측정 가능하다면 (이를 수학에서 $A \in \mathscr{F}$라고 표현하는 겁니다.) 그의 여집합 또한 측정할 수 있어야 할 것입니다. 마지막으로 측정 가능한 집합들을 무한히 합친다해도 그 역시 측정 가능할 것입니다. 이를 단순히 수학이라는 언어로 기술한 것에 지나지 않습니다. 또한 수학의 함축성으로 인해 정의를 대폭 줄일 수 있었습니다. 위의 정의에서 i), ii) 정의를 이용하면 전체 집합 역시 측정가능하다는 것을 알 수 있고 ii), iii) 을 이용하면 합집합들 뿐 아니라 교집합들 역시 측정 가능하다는 사실을 알 수 있습니다.
대수는 일종의 규칙이라 하였는데, 따라서 대수가 있는 공간과 아닌 공간은 구별되어야 합니다. 따라서 우리는 이를 다음과 같이 수학적으로 정의합니다.

Definition 2 - Measurable Space.

Let S be a set, and let $\mathscr{F}$ be a σ-algebra of subsets of S. Then $(S, \mathscr{F})$ is called a measurable space. The elements of $\mathscr{F}$ are called measurable sets.

이제 대략적인 Notation은 정리했으니 조금 더 생각을 해봅시다. 어떤 집합 S에 대하여 가장 작은 σ-algebra와 가장 큰 σ-algebra는 뭘까요? 일단, 정의에 의해 공집합은 무조건 측정가능하고 두 번째 정의에 의해 전체집합도 측정가능하므로 가장 작은 시그마대수는 $\mathscr{F} = \{\emptyset, S\}$가 됩니다. 또한 시그마 대수는 어떤 집합의 부분집합의 모임이기 때문에 모든 부분집합의 집합(멱집합; Power set)이 가장 큰 시그마 대수가 될 것입니다. 이는 $\mathscr{F} = \mathscr{P}(S)$로 표기합니다.
고등학교 시절에 부분집합을 배울 때, 어떤 집합을 포함하는 부분집합의 개수는? 이라는 문제를 본 적이 있을 겁니다. 되게 의미 없는 일 같지만 수학에서는 어떤 집합을 포함하고 있는 지의 여부가 상당히 중요합니다. 시그마 대수에서도 마찬가지인데 어떤 집합을 반드시 포함하고 있는 시그마 대수를 "그 집합에 의하여 발생한 시그마 대수이다" 라고 정의할 것입니다. 수학적 정의는 다음과 같습니다.

Definition 3.

Let S be a set and G be a family of subsets of S. The smallest σ-algebra which contains G is called generated σ-algebra with respect to G denoted by $\sigma(G)$.

이제 이 정의를 아주 유명한 규칙 공간에 이용해보겠습니다. 바로 위상공간(Topological Space)에 말이죠.

Definition 4 - Borel σ-algebra.

The Borel or topological σ-algebra $\mathscr{B}$ of a topological space $(S, \mathscr{T})$ is the σ-algebra generated by $\mathscr{T}$.

이 정의를 이해하기 위하여 굳이 위상수학까지 공부하지 않아도 됩니다. 우리는 앞으로 관측 공간인 $\mathbb{R}^d$에서의 대수만 사용할 것이기 때문이죠. 간단히 말하자면 n차원 실수공간의 위상은 n차원 직육면체로 표현됩니다. 1차원에서는 직선, 2차원에서는 직사각형 이런 식이죠. 따라서 위의 정의에 따르면 이런 직사각형들이 모두 측정가능하다면 그 공간을 일컬어 Borel σ-algebra 가 존재하는 공간이라 합니다.
지금까지는 공간, 집합에 대해서만 다뤘는데 우리의 가장 중요한 것은 그 공간에 작용하는 함수입니다. 위상수학이나 해석학을 했다면 다음의 정의는 아주 쉽게 다가올 것입니다. 만일, 하지 않았더라도 받아들이면 됩니다.

Definition 5 - Measurable function.

Let $(S, \mathscr{F})$, $(S', \mathscr{F}')$ be measurable spaces and $f : S \to S'$. If $\forall X \in \mathscr{F}'$, $f^{-1}(X) \in \mathscr{F}$ then f is called measurable function.

이 정의를 어떻게 사용하는지 간단한 예시를 들어 설명해보겠습니다.

Definition 6 - Indicator Function.

The indicator function of a subset A of a set X is a function

$$I_A(x) : X \to \{0,1\}$$

defined as

$$I_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Example 3 - Indicator function is measurable.

Let $A \in \mathscr{F}$ then I_A is measurable function.

위의 예시의 증명은 아주 간단하므로 생략하겠습니다. 꼭 한 번 해보고 넘어가면 이해가 잘 될테니 꼭 해보시길 바랍니다.